# dbSNP: a database of single nucleotide polymorphisms

**Elizabeth M. Smigielski, Karl Sirotkin[1], Minghong Ward[1] and Stephen T. Sherry[1,*]**

National Library of Medicine and [1]National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, MD 20894, USA

Received September 2, 1999; Revised and Accepted November 1, 1999

## ABSTRACT

**In response to a need for a general catalog of genome variation to address the large-scale sampling designs required by association studies, gene mapping and evolutionary biology, the National Cancer for Biotechnology Information (NCBI) has established the dbSNP database. Submissions to dbSNP will be integrated with other sources of information at NCBI such as GenBank, PubMed, LocusLink and the Human Genome Project data. The complete contents of dbSNP are available to the public at website: http://www.ncbi.nlm.nih.gov/SNP . Submitted SNPs can also be downloaded via anonymous FTP at ftp://ncbi.nlm.nih.gov/snp/**

## BACKGROUND

A key aspect of research in genetics is the association of sequence variation with heritable phenotypes. Occurring roughly every 500–1000 base pairs, single nucleotide polymorphisms (SNPs) are among the most common genetic variation. There is currently great interest in SNP discovery since a dense catalog of SNPs is expected to facilitate large-scale studies in association genetics (1), functional and pharmaco-genomics (2), population genetics and evolutionary biology (3), and positional cloning and physical mapping (4). To serve this need for such a general catalog, the National Center for Biotechnology Information (NCBI) established the Single Nucleotide Polymorphism database (5) (http://www.ncbi.nlm.nih.gov/SNP ) in collaboration with the National Human Genome Research Institute (NHGRI).

Since its inception in September 1998, the dbSNP database has served as a central, public repository for genetic variation. Once such variations are identified and catalogued in the database, additional laboratories can use the sequence information around the polymorphism and the specific experimental conditions for further research applications. As with all NCBI resources, the data within dbSNP is available freely and in a variety of forms.

## SCOPE

dbSNP currently classifies nucleotide sequence variations with the following types and percentage composition of the database: single nucleotide substitutions (97.8%), microsatellite repeats (0.1%) and small insertion/deletion polymorphisms (2.1%). There is no requirement or assumption about minimum allele frequencies or functional neutrality for the polymorphisms in the database. Thus, the scope of dbSNP includes disease-causing clinical mutations as well as neutral polymorphisms. In addition to the record identifiers assigned by both the submitter and NCBI, dbSNP entries record the sequence information around the polymorphism, the specific experimental conditions necessary to perform an experiment, descriptions of the population containing the variation, and frequency information by population or individual genotype.

The current level of activity in the discovery of general sequence variation suggests that SNP markers with unknown selective effects will be the majority of submitted records. Although most submissions are currently for *Homo sapiens*, dbSNP already has submissions for *Mus musculus*, and in general the database can accept variation information from any species and from any part of a particular genome. dbSNP is currently integrated with other large public variation databases such as the NCI CGAP-GAI database of EST-derived SNPs (6) and the TSC (The SNP Consortium, Ltd) variation initiative (7). Links to these, and other future public databases are established by the LinkOut scheme discussed below.

## UTILITY

dbSNP links variations (polymorphisms and clinical mutations) to other NCBI sequence resources via BLAST and E-PCR analysis of the flanking sequence that immediately surrounds the variation. Links to the literature databases are made with the citation information provided at submission time. This integration process makes dbSNP part of the NCBI 'discovery space' as illustrated in Figure 1. In this model, dbSNP serves dual roles as both a 'first point of entry' into the resource network for query and retrieval of specific variation records, and as an information server for searches that start in other resources such as GenBank, PubMed, LocusLink or the genome sequence databases.

As the final results of various genome projects accumulate, it is intended that all variations will be associated with a nucleotide sequence record and/or physical map contig. In the soon approaching post-sequencing phase of the human genome project, annotation of the sequence with features such as new genes or regulatory regions will provide new functional contexts for currently 'anonymous' variations that have been found on random sequence. As records appear for these new genes, links to dbSNP variations will be automatically annotated on the appropriate Reference Sequence or UniGene cluster. Resource integration of variation records extends beyond

[*] To whom correspondence should be addressed. Tel: +1 301 435 7799; Fax: +1 301 480 9241; Email: sherry@ncbi.nlm.nih.gov

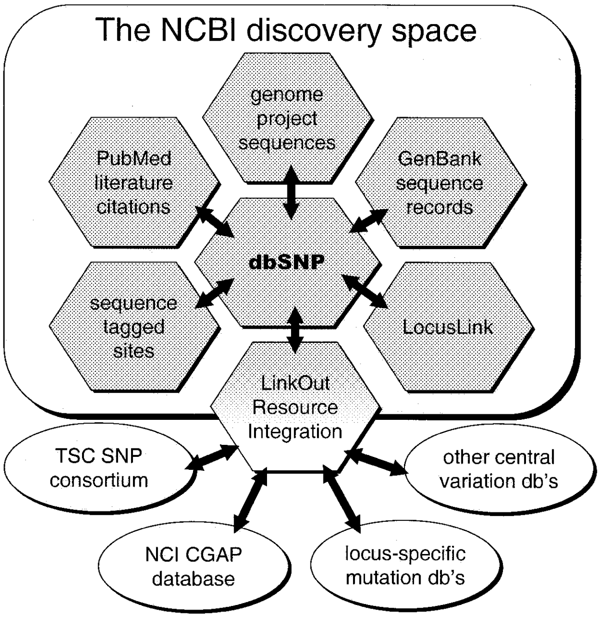

**Figure 1.** Records in dbSNP are cross-annotated within other internal information resources such as PubMed, genome project sequences, GenBank records, the LocusLink nomenclature/sequence database and the dbSTS database of sequence tagged sites. Users may query dbSNP directly, or start a search in any part of the NCBI discovery space to construct a set of dbSNP records that satisfies their search conditions. Records are also integrated with external information resources through hypertext URLs that dbSNP users can follow to explore the detailed information that is beyond the scope of dbSNP curation.

NCBI by use of 'LinkOut URLs' that refer to external databases with further information about the variation. This integration is important when one considers the general task of effectively annotating a complete genome for variation and its consequences for the organism (Fig. 2). NCBI has adopted the model in which variations are cataloged in dbSNP while functional descriptions of the local sequence region are noted as GenBank, dbSTS, Reference Sequence, LocusLink or UniGene records.

Since genes and their component nucleotides are potentially involved in multiple pathways and hence multiple downstream phenotypes, NCBI does not annotate the detailed biochemical or phenotypic consequences of variation directly on the sequence. Rather, links are maintained in dbSNP to external databases that each characterize particular axes of phenotypic variation, in much the same way that LocusLink maintains a current set of sequence accessions and nomenclature information for genes. In this fashion, dbSNP records can be linked to more complete descriptions of individual variations in locus-specific mutation databases. A federation of such databases can be found online at http://ariel.ucs.unimelb.edu.au:80/~cotton/guide1.htm

We are designing dbSNP to facilitate searches along four major axes of information: (i) sequence location, (ii) function, (iii) cross-species homology and (iv) degree of heterozygosity (degree of population variation). By setting thresholds of inclusion on one or more of these axes, users can extract the subset of records that are best suited to their research needs.

## SCALE

As of this writing, dbSNP contains 18 250 submissions from 55 groups describing variation in two species (human and mouse). Submissions can be divided into four general categories with the following percentages of the total database size: (i) mined from EST databases, e.g. (6), 57%; (ii) private investigator/corporate experimental results, 33%; (iii) early results of the NHGRI SNP discovery RFA, 9%; and (iv) SNP mining from the Human Genome Project sequences, 1%. Public and private initiatives to discover new SNPs in humans should provide an additional 350 000 submissions over the next 2–3 years (7,8). Additional submissions are expected from the human genome sequencing project where SNPs are being mined from the BAC end sequence alignments that are being used to construct the sequence contigs. This data mining project should produce another 300 000 SNPs within the year. These SNPs will be clustered within an estimated 30 000 sequence overlap regions that are dispersed throughout the genome. Collectively, these major initiatives in SNP discovery should report ~650 000 SNPs within 2 years.

## SUBMISSION

Submissions are welcomed from all sources, public and private. These groups are working on a variety of aspects of new SNP discovery, new technologies for SNP detection and rapid SNP genotyping in large samples. Data can be submitted directly to NCBI via instructions on the dbSNP 'How to Submit' Web page (http://www.ncbi.nlm.nih.gov/SNP/get_html.cgi?whichHtml=how_to_submit ). Required submission information includes the observed alleles at a particular locus, the flanking sequence that surrounds the mutation, the experimental methods used, and a pointer to a companion STS or GenBank record. Each individual laboratory is assigned a 'handle' to serve as a unique identifier, which will allow submissions to be associated with a specific laboratory. NCBI will also assign SNP accessioning, i.e., ss#, to each submitted SNP. A reference identifier will also be assigned to each unique SNP in an organism reference genome. These will be used to map the SNP to external resources or databases, including other NCBI databases.

## SEARCHING

dbSNP can be searched directly or via other NCBI resources that comprise the NCBI discovery space as illustrated in Figure 1. Direct searching can be done by submitter handle (laboratory), new batches of submissions, identification method used, population type studied, publication title, level of population variation or STS mapping information. As an integrated part of NCBI, the contents of dbSNP are cross-linked to records in other information resources such as GenBank, LocusLink, the human genome sequence and PubMed. The result sets from queries in any of these resources will point the user back to the relevant records in dbSNP.

*BLAST.* dbSNP can be searched with the standard BLAST algorithm that will compare a user-submitted sequence against all flanking sequence records in dbSNP. The BLAST service is

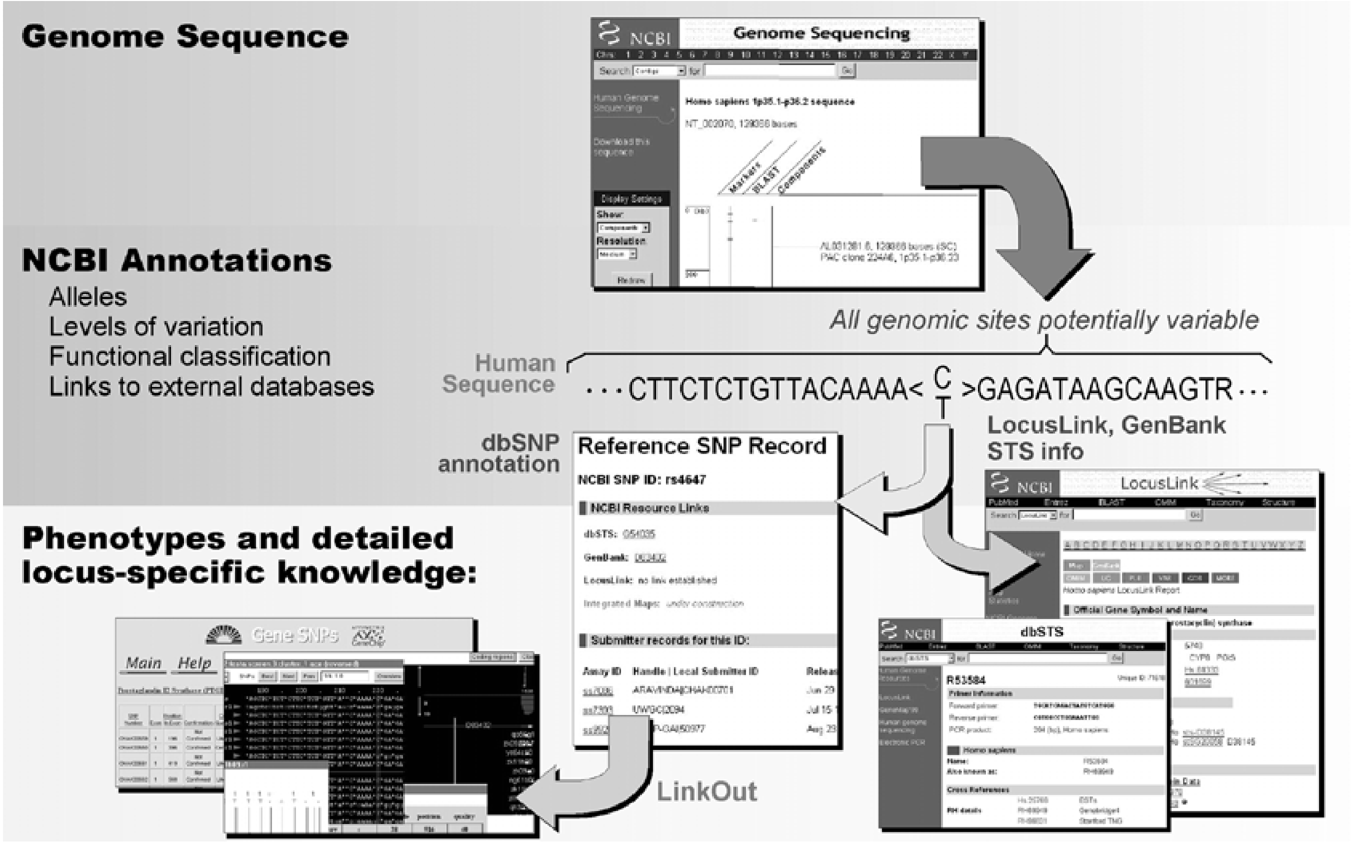
Genome Sequence
NCBI Genome Sequencing
Home sapiens 1p35.1-p36.2 sequence
NT_002070, 129366 bases
Markers
BLAST
Components
NCBI Annotations
Alleles
Levels of variation
Functional classification
Links to external databases
All genomic sites potentially variable
Human Sequence
· · · CTTCTCTGTTACAAAA< C T >GAGATAAGCAAGTR · · ·
dbSNP annotation
Reference SNP Record
NCBI SNP ID: rs4647
NCBI Resource Links
dbSTS:
GenBank:
LocusLink: no link established
Integrated Maps: under construction
Submitter records for this ID:
Assay ID
Handle | Local Submitter ID
LocusLink, GenBank STS info
NCBI LocusLink
NCBI dbSTS
R53584
Phenotypes and detailed locus-specific knowledge:
Gene SNPs
Main
Help
LinkOut

**Figure 2.** Annotation of the multiple biochemical or phenotypic consequences of individual variations in genomic sequence is accomplished through (i) the annotation of a single dbSNP record on the genome sequence to indicate the presence and extent of variation, and (ii) the maintenance of a list of accession links and URLs within the dbSNP record to other information resources. In this way a single variation can be easily represented in multiple biochemical pathways or phenotypic backgrounds.

provided on the dbSNP homepage, rather than the general NCBI BLAST page.

*LocusLink.* dbSNP can also be queried by integrating it with other NCBI resources. Via LocusLink, queries can be done by gene name or nomenclature association. Query results from the LocusLink database will show a purple 'V' button in SNP records which have been mapped to the gene. Clicking on the this button will lead to a list of the reference SNP records for any gene in the LocusLink database.

*Entrez.* The sidebar of the 'graphical view' has a 'SNP' link to dbSNP records linked to the GenBank accession number.

*Genome sequence.* The contig view can be set to show 'variations' in addition to STS 'markers' and sequence 'components'.

## DOCUMENTATION

The database is updated after each new data submission. A regularly updated database summary documents the number of SNPs identified, submitters, publications cited, and methods and populations defined. Complete submission guidelines are available on the dbSNP website. A FAQ page lists frequently asked questions derived from user inquiries.

## FUTURE PLANS

dbSNP is currently being integrated to GeneMap99 and the integrated physical maps that are being constructed at NCBI. When integration is completed, the maps will be browsable for SNP content in user specified regions of the map.

In addition to this integration with other NCBI resources, enhancements to the interface will improve searching and data submission. Expanded query facilities and graphical user interfaces will permit structured queries and batch retrieval of results. Online Web data submission will complement the established batch submission process.

dbSNP is a relatively new database. Although many small contributors submit data, the majority of data is expected from a few large research projects. For this reason, dbSNP is expected to grow rapidly over the next few years. Data exchange with other public variation and mutation databases and the extension of the database to support haplotype data objects will also increase the amount of data in dbSNP and enhance its utility.

## REFERENCING dbSNP

We suggest that dbSNP be referenced as follows: Sherry,S.T., Ward,M. and Sirotkin,K. (1999) dbSNP—Database for Single

Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. *Genome Res*., **9**, 677–679.

We suggest that the abbreviation dbSNP be used for this database.

## ADDRESSES

For assistance in using dbSNP, please write to info@ncbi.nlm.nih.gov . For other questions regarding dbSNP, please contact Stephen Sherry at sherry@ncbi.nlm.nih.gov. Mail may be addressed to Steve Sherry, National Center for Biotechnology Information, National Library of Medicine, Building 38A, Room 8N805, Bethesda, MD 20894, USA. Tel: +1 301 435 7799; Fax: +1 301 480 9241.

## ACKNOWLEDGEMENTS

This research was supported in part by an appointment to the NLM Associate Fellowship Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.

## REFERENCES

1. Kruglyak,L. (1999) *Nature Genet.*, **22**, 139–144.
2. Carulli,J.P., Artinger,M., Swain,P.M., Root,C.D., Chee,L., Tulig,C., Guerin,J., Osborne,M., Stein,G., Lian,J. and Lomedico,P.T. (1998) *J. Cell Biochem.*, **30–31** (Suppl.), 286–296.
3. Cavalli-Sforza,L.L. (1998) *Trends Genet.*, **14**, 60–65.
4. Collins,F.S. (1999) *N. Engl. J. Med.*, **341**, 28–37.
5. Sherry,S.T., Ward,M. and Sirotkin,K. (1999) *Genome Res.*, **9**, 677–679.
6. Buetow,K.H., Edmonson,M.N. and Cassidy,A.B. (1999) *Nature Genet.*, **21**, 323–325.
7. Masood,E. (1999) *Nature*, **398**, 545–546.
8. National Institutes of Health (US) Office of Extramural Research. *Methods for Discovering and Scoring Single Nucleotide Polymorphisms*. NIH guide for grants and contracts [computer file]/NIH.RFA: HG-98-001. 1-9-98.